International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 5 Issue 4, May-June 2021 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Discovering Student Dropout Prediction through Deep Learning

Shashikant Karale, Rajani Pawar, Sharvari Pawar, Poonam Sonkamble

Student, Padmabhushan Vasantdada Patil Institute of Technology, Pune, Maharashtra, India

ABSTRACT

There have been increased incidences of dropout that have been noticed in the universities in the recent years. These increased reports have been instrumental in introducing the graduation rate of the course completion rate for majority of universities all over the globe. Dropouts are highly undesirable and are an indication of some underlying inconsistencies that have been plaguing the course since a long time. Therefore, an effective system for the purpose of prediction of the dropout rate is the need of the hour. To reach these goals this research article has utilized machine learning approaches. The proposed methodology utilizes the K Nearest Neighbor, Fuzzy Artificial Neural Network and Decision Tree. This approach has been illustrated in utmost detail in this research article, highlighting the execution of the various important modules of the methodology. The experimentation has been performed to achieve the performance of the approach which has yielded highly accurate results.

KEYWORDS: *Fuzzy Artificial Neural Networks, K Nearest Neighbor, Decision Tree, Online Courses*

How to cite this paper: Shashikant Karale | Rajani Pawar | Sharvari Pawar | Poonam Sonkamble "Discovering Student Dropout Prediction through Deep Learning" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-5 | Issue-4, June 2021, pp.1549-1553, URL: www.ijtsrd.com/papers/ijtsrd43700.pdf

IJTSRD43700

INTRODUCTION

Education is an essential and one of the most integral forms of development for a human being. The human being is one of the most intellectually capable organisms on this planet. With the highest brain to body ratio, a lot of activities and other important virtues are the direct result of an increased intellect. As the case with having an intellectual power it needs to be useful and should be honed effectively in a particular way. This honing of the internet is one of the most crucial aspects of growing up. As a child grows up over the years more and more knowledge is accumulated and the child learn new ways and techniques to become self-sustainable and improve his conscience.

Education plays an important role in improving the conscience of the individual significantly. This improvement it is done through providing and effective education to the children starting from a very young age. This development is essential for the overall health mentally and physically of the child. As the child grows learn more about the world around him and is educated by the parents and at school. School is an important part of a child's life which teaches him a lot of things that would help them later in life and currently. The schools are a heaven of knowledge which imparts valuable education to the children effectively improving their lives over all.

After the education provided by the schools these children then effectively are introduced to various different concepts and Fields. And the children utilize these fields and guess their interest to pursue education in those particular interested fields. This in depth education is entrusted form by the Universities and colleges. The University provides education effectively through the use of lecturers and classes. The interested students pursue the relevant course for a period of time stipulated by the University to get proficient in their field. This is essential for the student as well as the educational institution to provide the valuable education to the students. As universities are designed in a capitalistic model they need to make profits to keep providing education to large number of students.

But there has been noticed that there are increase incidences of student dropout from the courses over the past few years. These dropouts are highly problematic and can be effectively detrimental to the university as well as students. The increased numbers of dropouts are basically an indication of some irregularities that are forcing the students to dropout. This process can be ameliorated if the intentions of the students to dropout are effectively visible in time. Therefore there is a need for an effective system for the purpose of predicting the student dropout rate in a particular course. These predictions can be highly valuable and decreasing the dropouts and helping the students complete their course or degree without any hiccups.

There have been number of related words that have been analyzed in this research article. These researches have been performed to provide resolution for the increasing rates of dropout that are noticed in universities all over the globe. There has been increased interest of the researchers

introducing this problem and finding the root cause of the increased dropout rates. The findings have been effective in understanding that the prediction of a dropout rate if determine for a student it can be effectively useful in addressing the problems and the concerns of the student. This information is valuable in reducing the rate of dropouts for the students and improving the rate of successful completion for the university. The prediction has also been useful in understanding the inconsistencies in the course material which can be rectified in time before a lot of student's dropout. Therefore the related words analyzed in this approach have been effective in the designing of our technique for the purpose of dropout prediction through the implementation of machine learning approaches. Literature survey paper provides a thorough analysis of the student dropout rate prediction and realizes an effective machine learning implementation for improving the dropout prediction.

In this paper, section 2 is dedicated for the literature review of past work, Section 3 describes the details of the developmental procedure of the model. Section 4 evaluates the results through some experiments and finally section 5 concludes this research article with the traces of the future scope.

Literature Review

T. Hasbun narrates in Organization for Economic Co-operation and Development countries student dropout is one of the salient problems. [1] Thus for dropout prediction, Educational Data Mining (EDM) analysts are studying institutional interventions through students. In the proposed paper the authors have given importance to extracurricular activities from science degrees engineering and business to predict dropout. The main aim of this paper is to prove that how extracurricular activities can be suitable factors for dropout prediction. The results of the proposed technique have shown better performance.

Di Sun states in the last few years there has been a rapid growth in open online courses (MOOCs) but they're a serious problem is arising by the MOOC researchers and providers is dropout prediction. [2] The paper presents a solution for students getting dropouts just by analyzing how much syllabus can be completed by the students in a course. Thus for predicting the student dropout recurrent neural network (RNN) and learning resource representation layer is used thus the accuracy of the proposed papers is higher than the traditional machine learning methodology.

J. Canas aims to offer an easy and explainable plan of action to recognize dropout-prone and fail-prone students.[3] For practitioners and researchers students who do not complete there higher education or they interrupt their study in between is one of the major concern. The early prediction of the student's dropout may help teachers to reduce dropout rates and also suggest new learning materials for the students from failing or not completing the course. In the proposed paper the researcher has used tree-based classification models to predict dropout-prone and fail-prone students.

M. Mustafa describes for developing countries the national and international loss of student's dropout prediction is one of the major concerns. The main aim of this paper is to implement an effective dropout prediction system for institutes and colleges.[4] To classify the successful from unsuccessful students they implemented square test methodology on factors such as financial condition, dropping year, and gender. They have used feature selection, logistic regression, and neural networks to data using data mining techniques Classification and Regression Tree (CART) and CHAm tree.

L. Wang explains the emerging wave of artificial intelligence and network information platform technology for the Massive Open Online Courses (MOOC). MOOC is a new classroom model that is greatly hit on the traditional education model.[5] In the proposed paper they implemented a prediction model by using a time-controlled Long Short-Term Memory neural network (ELSTM).[5]Tasks such as learning evaluation, talent training, curriculum development, teaching mode have greater benefits. Some many shortcomings and problems need to be solved in MOOC. The experimental results of the proposed model are higher which shows the effectiveness of the proposed paper.

N. Shiratori states since the 1990s a serious problem has been are rising in Japan that is dropout rate has been a major point of interest by Yomiuri Shimbun. [6] As the semester progresses students lose the interest in course. In preliminary dropout, status distributes the dropout-related state of students by using the logistic regression model. Thus the proposed paper implements model to divide students to examine the issue of student dropout rates. Thus the proposed research students with low motivation in studies so faculty member to know the reasons for this low motivation an early interview.

D. Ktoridou narrates that almost half of engineering students change their field due to teaching and advising insufficiency, the complexity of the engineering curriculum.[7] There are six personality types enterprising, conventional, artistic, social, and realistic. The proposed paper shows the similarity between engineering undergraduate students with personality types. This research is done with 113 undergraduate students the main reason for leaving engineering was poor teaching and advising. The paper provides positive effects on both students and Engineering departments

R. Pereira presents the student dropout prediction thus by using the data mining techniques disciplinary and institutional data of students from the University of Nariño. [8] By using the classification techniques such as decision trees Socioeconomic and academic student dropout profiles were generated. The proposed model can help to develop policies and strategies for the students to improve the activities for studies. In recent years there have been many types of research made in the field of education by using Data mining techniques.

A. Ortigosa states the main aim of the proposed paper is to put a stop to students from leaving the university.[9] The proposed paper collected 11,000 students' data from the five years by using the C5.0 algorithm which is one of the successful predictive models and SPA i.e Sistema de Prediccioen de Abandono dropout prediction system in Spanish. The SPA model was taken into use in 2017 and is currently in use. Thus by using this model 117,000 risk scores to detect the dropout risk of more than 5700 students it to support student dropout risk prevention.

K. Murakami performs machine learning techniques to predict dropout rates of current and graduate students by collecting their data thus the main aim of the paper is to

dropout prevention. [10] The university arranges and collects data of students to provide enrollment management. They focus on the student dropout and by using data from both current, graduate, and dropout students. Thus by using the machine learning algorithm while predicting these dropout rates it is important to investigate unknown items regarding EMIR.

K. Santos explains for university student's dropout prevention by using various educational data mining techniques. [11] The algorithm such as RF and Decision Trees some algorithms are not successful to get high outcomes due to the large data and a parameter tuning step. Thus the proposed paper uses the different supervised learning techniques such as Decision Trees, K-nearest Neighbor, Support Vector Machines, Naive Bayes, Neural Networks, and Random Forests to calculate the prediction models for reducing school dropout.

R. Lottering describes in the field of education in South Africa an advanced development has been seen in this development dropout rate is noticed. [12] As in higher education in South Africa student numbers grew by 32.8% from 2006 to 2015 and the dropout rate of 17.1% has been noticed and it can be doubled in the next ten years. Thus proposed systems use data mining techniques and predict dropout goals. The researches of the proposed papers show a positive approach to the proposed model.

PROPOSED METHODOLOGY

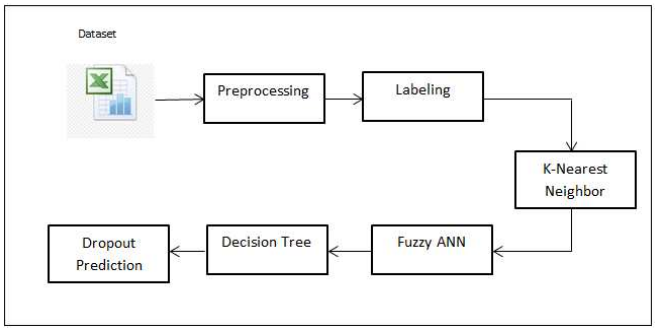

Figure 1: Overview of the Proposed Idea

The proposed methodology of the learning outcomes of massive online courses is depicted in figure 1 and it is explained through following steps

Step 1: *Preprocessing and Labeling* – The proposed system uses the dataset of massive online courses for the purpose of experimentation. The dataset is downloaded from the URL : https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/29779.

Once Testing dataset of the user is received, it is statically stored for the instance learning outcomes. And then proposed model uses the Training set where it reads and stored in a double dimension list, which is subjected to preprocessing thereafter. This double dimension list of dataset belongs to training data is divided into certified and non-certified set based on the Boolean value present in the column of the dataset.

Then this dataset is used to label them with the unique integer numerical value using hash set functions. After this all the required rows of the preprocessed list are assigned with the integer numerical value which is then subjected to clustering process in the next step of the model.

Step 2: K Nearest Neighbor – The preprocessed list attained in the previous step is utilized as an input in this step of the procedure. This step of the student dropout detection performs the K Nearest Neighbor clustering. This is performed to classify the dataset given as an input into semantic groups to achieve accurate detection of the student dropout in massive online courses. This procedure is performed through the steps given below.

Distance Evaluation – The Euclidean Distance formula given in the equation 1 below is being used for the purpose of evaluation of the distance between the training and testing dataset attributes. The dataset is provided in the form of a list that is given as an input to this step of the procedure. The data stored in the rows of this list is used for the purpose of calculating the distance from the labeled testing dataset, and stored as the row distance at the end of the list. This is performed for all the entries in the list iteratively.

$$ED = \sqrt{\sum(ATi - ATj)^2} \quad \text{_____ (1)}$$

Where,

ED=Euclidian Distance

A_{Ti}=Attribute at index i

A_{Tj}= Attribute at index j

Centroid Estimation – The list with the data and the respective row distances calculated in the previous step are used in this step as an input. The list is sorted in the ascending order of the Row distance and subjected to random data point selection. These selected data points are K in number and achieving the centroid. These data points are used to determine the row distance of the selected index. The obtained row distance is used further to achieve the boundary of the clusters.

The row distances achieved previously are used to form the average row distance of the entire list. These average row distances along with the extracted row distance of the centroid are used to determine the boundaries of the clusters in the cluster formation procedure given below.

Cluster Formation – The centroids and the average row distance acquired previously are provided as an input to this step for cluster formation. This step then calculates the boundaries of the clusters by addition and subtraction of the average row distance and the centroid row distance to attain the maximum and minimum values respectively. The data is then subjected to these boundaries to form the clusters which are aggregated into a cluster list. This cluster list sorted into a descending order and the top 3 clusters are provided to the next step as an input.

Step 3: Fuzzy Artificial Neural Network – This is one of the most important steps in the proposed methodology wherein a double dimensional list is utilized to effectively store the attributes of the student details as well as the clusters of these values. The fuzzy classification is achieved on this list to effectively segregate the clusters and effectively identify the difference between the maximum and the minimum cluster values. The difference achieved is then effectively divided into 5 equal parts to achieve effective fuzzy classification labels. These labels correspond to the fuzzy crisp values that are segregated as very high medium-low and very low. The student ID and the score sum of different activities are stored in a double list to generate the fuzzy crisp ranges as mentioned in the below algorithm1.

ALGORITHM 1: Fuzzy Crisp Value generation

//Input : Student Score list S_{LST}

//Output: FC_{LST} (Fuzzy Crisp List)

fuuzyCrispValueGeneration(S_{LST})

1: Start
2: Set min=0.5, max=0.5
3: ***for*** i=0 to size of S_{LST}
4: $TMP_{LST} = S_{LST[i]}$ [TMP_{LST} = Temporary Set]
5: $PR_{SCRE} = TMP_{LST[1]}$ [Score]
6: **IF** (PR_{SCRE} < min)
7: min= PR_{SCRE}
8: **IF**(PR_{SCRE} >max)
9: max= PR_{SCRE}
10: ***end for***
11: RANGE1=0 , RANGE2=0 ,R_{LIST}= Ø [Rule List]
12: DF=(max-min)/5 [DF= Diffrence Distance]
13: ***for*** i=0 to 5
14: RANGE1=min
15: RANGE2=RANGE1+DF
16: min=RANGE2
17: $T_{LST[0]}$= RANGE1 [T_{LST}= Temporary List]
18: $T_{LST[1]}$= RANGE2
19: $FC_{LST} = FC_{LST} + T_{LST}$
20: ***end for***
21: return FC_{LST}
22: **Stop**

The neurons for the artificial neural network are then classified according to the given rules above depending on the classified certified set with labels provided such as very high, high, medium, low and very low. These neurons are then subjected to affective evaluation of the hidden layer and the resultant probability achieved effectively determines the dropout rate for the desired level according to the user

Step 4: Decision Tree – The probability values achieved from the previous step of fuzzy artificial neural networks is taken as an input to this step of the methodology. The Decision tree approach is one of the most powerful and highly accurate classification algorithms that completely categorize the data and segregate is effectively. The decision tree approach utilizes the if-then rules to accomplish its goals of classification based on the threshold values applied for execution. The probability scores related to the student dropout scenario have been subjected to the classification to achieve highly accurate classification of the dropout predictions. This achieves the student dropout results that are displayed to the user through the interactive user interface.

RESULTS AND DISCUSSIONS

The proposed system for student dropout prediction through the use of Fuzzy Artificial Neural Networks has been achieved in the Java programming language. The laptop utilized for the development of this prediction methodology is equipped with an Intel core i3 processor which is supplemented by 500 GB of storage and 6 GB of RAM.

For the purpose of achieving the performance metrics of the proposed methodology the precision and recall parameters have been assigned. The extraction of the performance of this approach is necessary to quantify the accuracy of the student dropout methodology which is being developed through Fuzzy ANN and decision tree algorithms. The measurement of the accuracy of the approach will be useful in determining if these algorithms have been accurately implemented which will be evident through the scores achieved by the precision and recall metrics.

Performance Evaluation based on Precision and Recall

The performance metrics through precision and recall have been assessed through the use of intensive experimentation that has been performed on the proposed methodology. These performance parameters are useful in determining the real accuracy of the evaluation and the actual performance of the methodology. The student dropout detection is highly useful and can only be viable with a respectable accuracy of the identification through the methodology prescribed in this research work.

The in-depth information regarding the performance of the approach has been achieved through the precision and recall parameters where each of these parameters is useful in identifying a different aspect of performance of the methodology. The precision parameter is useful for determining the relative performance of the student dropout prediction on the other hand recall extracts the absolute accuracy for the student dropout prediction.

The precision for this methodology has been conceived as the division of the number of accurate student dropout predictions by the total number of expected student dropout predictions. The recall parameter is on the other hand the absolute accuracy which is derived by the ratio of total number of student dropout prediction versus the total number of expected student dropout prediction.

This has been mathematically conveyed through the equations 4 and 5 given below.

A = The number of correctly predicted student dropout

B= The number of incorrectly predicted student dropout

C = The number of student dropout not predicted

So, precision can be defined as

Precision = (A / (A+ B)) *100

Recall = (A / (A+ C)) *100

The in-depth experimentation done on the prescribed approach for the measurement of precision and recall parameters has been listed in the table 1 below.

No. of expected Predictions	The number of correctly predicted Student Dropouts (A)	The number of incorrectly predicted Student Dropouts (B)	The number of Student Dropouts not predicted (C)	Precision	Recall
38	27	5	6	84.375	81.81818182
67	48	9	10	84.21052632	82.75862069
87	63	10	14	86.30136986	81.81818182
119	89	19	11	82.40740741	89
139	98	20	21	83.05084746	82.35294118

Table 1: Precision and Recall Measurement Table for the performance of student dropout prediction

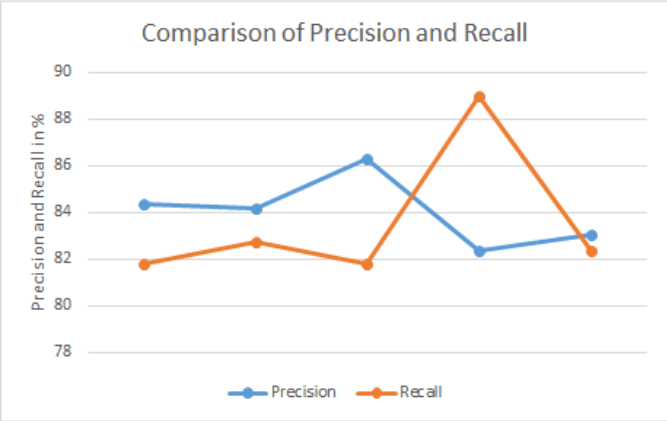

Figure 2: Comparison of Precision and Recall for the performance of student dropout prediction

The outcomes achieved in the table are effectively used to draw a plot graph for precision and recall parameters for graphical representation in the figure 2 above. As it is evident from the tabulated values and the outcome of the line graph, that the precision and recall parameters achieved for this methodology are well within respectable limits. The value of precision and recall as 84.06 and 83.54 are indicative of an accurate implementation of Fuzzy ANN and Decision Tree approaches that have been utilized for accurate student dropout prediction in the proposed methodology.

CONCLUSION

This research article has been effective in the analysis of predominant researchers on the topic of dropout prediction. The dropouts are highly undesirable occurrence that usually is a loss for the student as well as the university involved. The dropouts can hurt the universities negatively and could be one of the reasons for the increased dissatisfaction of the students with the course. These problems can only be solved through to the effective implementation of a dropout prediction approach. Therefore this research article has been effective in achieving methodology utilizing machine learning approaches for the purpose of dropout prediction. The input dataset is first effectively preprocessed and labeled before providing it to the K Nearest neighbor for the clustering. The KNN approach clusters the data using the certified labels. The generated clusters are provided to the Fuzzy Artificial Neural Networks as an input which processes the clusters using the fuzzy crisp values to achieve the predictions. These predictions are then effectively classified using the if-then rules of the Decision tree approach. The outcomes indicate the effective implementation of the student dropout methodology.

The future research directions can be focused on the realization of the student dropout methodology to be implemented in a web application for easier access to the required authorities.

References

[1] T. Hasbun, A. Araya and J. Villalon, "Extracurricular Activities as Dropout Prediction Factors in Higher Education Using Decision Trees," 2016 IEEE 16th International Conference on Advanced Learning Technologies (ICALT), Austin, TX, 2016, pp. 242-244, doi: 10.1109/ICALT.2016.66.

[2] D. Sun, Y. Mao, J. Du, P. Xu, Q. Zheng and H. Sun, "Deep Learning for Dropout Prediction in MOOCs," 2019 Eighth International Conference on Educational Innovation through Technology (EITT), Biloxi, MS, USA, 2019, pp. 87-90, doi: 10.1109/EITT.2019.00025.

[3] J. Figueroa-Cañas and T. Sancho-Vinuesa, "Early Prediction of Dropout and Final Exam Performance in an Online Statistics Course," in IEEE Revista Iberoamericana de Tecnologias del Aprendizaje, vol. 15, no. 2, pp. 86-94, May 2020, doi: 10.1109/RITA.2020.2987727.

[4] M. N. Mustafa, L. Chowdhury and M. S. Kamal, "Students dropout prediction for intelligent system from tertiary level in developing country," 2012 International Conference on Informatics, Electronics & Vision (ICIEV), Dhaka, 2012, pp. 113-118, doi: 10.1109/ICIEV.2012.6317441.

[5] L. Wang and H. Wang, "Learning Behavior Analysis and Dropout Rate Prediction Based on MOOCs Data," 2019 10th International Conference on Information Technology in Medicine and Education (ITME), Qingdao, China, 2019, pp. 419-423, doi: 10.1109/ITME.2019.00100.

[6] N. Shiratori, "Derivation of Student Patterns in a Preliminary Dropout State and Identification of Measures for Reducing Student Dropouts," 2018 7th International Congress on Advanced Applied Informatics (IIAI-AAI), Yonago, Japan, 2018, pp. 497-500, doi: 10.1109/IIAI-AAI.2018.00108.

[7] D. Ktoridou and E. Epaminonda, "Measuring the compatibility between engineering students' personality types and major of study: A first step towards preventing engineering education dropouts," 2014 IEEE Global Engineering Education Conference (EDUCON), Istanbul, 2014, pp. 192-195, doi: 10.1109/EDUCON.2014.6826089.

[8] R. Timaran Pereira and J. Caicedo Zambrano, "Application of Decision Trees for Detection of Student Dropout Profiles," 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), Cancun, 2017, pp. 528-531, doi: 10.1109/ICMLA.2017.0-107.

[9] A. Ortigosa, R. M. Carro, J. Bravo-Agapito, D. Lizcano, J. J. Alcolea and Ó. Blanco, "From Lab to Production: Lessons Learnt and Real-Life Challenges of an Early Student-Dropout Prevention System," in IEEE Transactions on Learning Technologies, vol. 12, no. 2, pp. 264-277, 1 April-June 2019, doi: 10.1109/TLT.2019.2911608.

[10] K. Murakami et al., "Predicting the Probability of Student Dropout through EMIR Using Data from Current and Graduate Students," 2018 7th International Congress on Advanced Applied Informatics (IIAI-AAI), Yonago, Japan, 2018, pp. 478-481, doi: 10.1109/IIAI-AAI.2018.00103.

[11] K. J. de O. Santos, A. G. Menezes, A. B. de Carvalho and C. A. E. Montesco, "Supervised Learning in the Context of Educational Data Mining to Avoid University Students Dropout," 2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT), Macei??, Brazil, 2019, pp. 207-208, doi: 10.1109/ICALT.2019.00068.

[12] R. Lottering, R. Hans and M. Lall, "A model for the identification of students at risk of dropout at a university of technology," 2020 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD), Durban, South Africa, 2020, pp. 1-8, doi: 10.1109/icABCD49160.2020.9183874.